User guide for EMI (v1)

June 27, 2013

1 Install

To install the program: sh install.sh
To see a quick help document: emi -help
To see a sample run: less test.sh

2 File format

The input file include pairwise IBD segments information, with 9 columns:
1) First sample identifier
2) First sample haplotype index (0 or 1)
3) Second sample identifier
4) Second sample haplotype index (0 or 1)
5) Starting genomic position (bp)
6) Ending genomic position (bp)
7) LOD score (eg. from beagle output, larger values indicate greater evidence for IBD)
8) Starting genomic position (cM)
9) Ending genomic position (cM)

For column 5-9, use 0 if the value is missing.

The output file indicate cluster starting position and end positions in column 2 and 3.

3 Command line arguments

```
./emi[input seg file] <parameters> [cluster output file]}
```

EMI (version 1) scans the genome along sliding windows and build multiple-IBD clusters based on pairwise IBD segments. The following are two example runs of the test data, with window size based on physical distance and genetic distance respectively.

```
./emi test.ibd -fam test.fam -wgt bp 2000000 6000000 -win 20000 bp test
./emi test.ibd -fam test.fam -wgt cM 1 7 -min 3 -win 0.15 cM -den 0.5 test
./emi test.ibd -fam test.fam -wgt 7th 3 40 -min 3 -win 0.15 cM -den 0.5 test
```

The input pairwise IBD segment file is test.ibd and the output file is test.tmp. A post-processing step is needed to remove the identical multiple-IBD clusters. An example run is as following:

```
less test.clst.tmp | awk -F '\t' -v OFS='\t' "{$1='clst'; print $0}"
| sort -k2n -k3n -k4n|uniq > test.clst
```

The required command line arguments include:
-help Print this output
-win [**STRING**] [**VAL1**]Sliding window size (STRING) and scale (VAL1 = bp or cM)
-wgt [**STRING**] [**VAL1**] [**VAL2**]Weight the input pairwise IBD (STRING) by length (in bp or cM) or value of 7th column of input file (larger values corresponding to higher confidence of pairwise IBD segments), the weighting score ranges from 0.8 to 1. [VAL1] has weight 0.8, segments with value smaller than [VAL1] are discarded, [VAL2] has weight 1, segments with value larger than [VAL2] has score 1.
-fam [**STRING**]Filename for list of sample identifies in PLINK format
-den [**VAL**]Minimum cluster density (default: 0.6)
-min [**VAL**]Minimum haplotype size (default: 3)